BUPT at TREC 2009: Entity Track

Zhanyi Wang, Dongxin Liu, Weiran Xu, Guang Chen, Jun Guo

Pattern Recognition and Intelligent System Lab,

Beijing University of Posts and Telecommunications, Beijing, China, 100876

wangzhanyi@gmail.com

Abstract. This report introduces the work of BUPT (PRIS) in Entity Track in TREC2009. The task and data are both new this year. In our work, an improved two-stage retrieval model is proposed according to the task. The first stage is document retrieval, in order to get the similarity of the query and documents. The second stage is to find the relationship between documents and entities. We also focus on entity extraction in the second stage and the final ranking.

1. Introduction

Entity track is a new one in TREC. The overall aim is to create a corpus for the evaluation of entity-related searches on Web data. [1] The task is related entity finding. It is motivated to a large degree by the expert finding task at the TREC Enterprise track. But they are different from many aspects, such as the definition of entity and task, entity type, primary homepage, even some specific method of IR and IE.

The task of related entity finding is defined as follow. Given an input entity, by its name and homepage, the type of the target entity, as well as the nature of their relation, described in free text, find related entities that are of target type, standing in the required relation to the input entity.

In our work, we improve a two-stage retrieval model according to the task. Moreover, we focus on entity extraction and ranking. The process of document retrieval is realized by indri toolkit [2], which is based on a combination of the language modeling and inference network retrieval frameworks.

The report is organized as follows. Section 2 introduces our search model briefly. Section 3 describes the process of related entity finding. Submitted runs shows in section 4 and section 5 gives the conclusion and Future work.

2. Search Model

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **NOV 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**
4. TITLE AND SUBTITLE **BUPT at TREC 2009: Entity Track**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Beijing University of Posts and Telecommunications,Pattern Recognition and Intelligent System Lab,Beijing, China, 100876,**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR’S ACRONYM(S)
		11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES **Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).**		
14. ABSTRACT **see report**		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

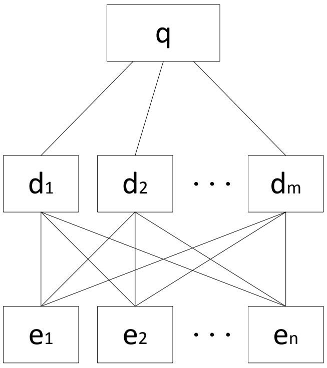


Figure 1. Search model

In order to finding entities and ranking support documents, a two-stage search model which is used in Enterprise track before is adopted. [3] Furthermore, we improve it according to the task. The structure is shown in Figure 1. Take documents (d_1, d_2,…,d_m) as a bridge between query(q) and entities(e_1, e_2,…,e_n). Then computing the score of query and entities transforms to two steps. The first step is to retrieve documents based a query. It is the same as general ad hoc retrieval. The second step is to find the relationship between documents and entities. The final score is got by combining previous results. The similarity of query and entity is:

$$Sim\left(q,e_j\right)=\sum_{i=1}^{m}Sim\left(q,d_i\right)\cdot Sim\left(d_i,e_j\right) \tag{1}$$

Here m is the number of relevant documents. d_i is the *i*st document and e_j is the *j*st entity.

The model is simple and original. In following sections, improvements for it will be elaborated, such as entity extraction and classification, computing the similarity of document and entity, also the final score which is used to rank entities.

3. Related Entity Finding

3.1 Overview

The system architecture is designed as Figure 2. First, we preprocess data which include many html tags. Meanwhile, some useful information (id, title, URL) is extracted. Second, we use indri toolkit to build two kinds of index. One is ordinary. The other makes a title field additionally. The index is the core in the system. Third, the represented or expanded query is sent to the indri, and then ranked documents and similarity of the query and documents return. Next, entities are extracted from the documents. We choose one type of the target entity and find ranked documents related to these entities by indri again, in order to get the similarity. Finally, similarities integrate to ranking score, also entities and support documents are given.

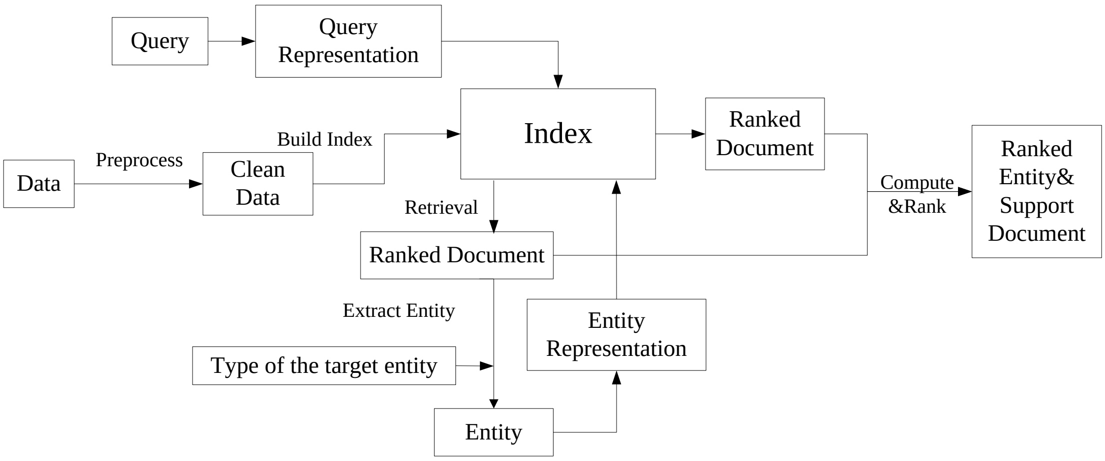
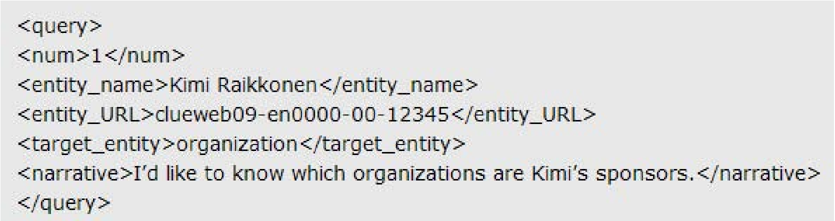


Figure 2. The system architecture

3.2 Document Retrieval Based Query

The process of document retrieval is ad hoc retrieval. We focus on three aspects, query representation, the special field in the text and the number of returned documents.

The type of a query is shown in figure 3. Besides the entity name, other three tags are also useful. An entity URL uniquely identifies an entity. It can be located to a document. The type of target entity is necessary. There are three types in the task, people, organization and product. Each query maps to a particular type of entity. So we only need to extract the just one. The narrative field is optional. However, we find that if we add some words into a query, the result will be better.



```
<query>
<num>1</num>
<entity_name>Kimi Raikkonen</entity_name>
<entity_URL>clueweb09-en0000-00-12345</entity_URL>
<target_entity>organization</target_entity>
<narrative>I'd like to know which organizations are Kimi's sponsors.</narrative>
</query>
```

Figure 3. An example query

An ordinary full-text index must be built. It can be taken as a baseline in the experiment. Moreover, we find the title is essence in each web page. So we use it as an additional field. In the preprocess stage, titles of web pages are extracted. Then we modify the parameter file and build an index that includes a title field.

We don’t need all relevant documents for the ones at the bottom of ranking list are almost irrelevant to the query. Their contribution to the final score can be negligible. In addition, the computational complexity increases with the scale of document. Taking into account the above factors, we control the number of relevant documents in the range of about 20 to 100.

3.3 Entity Extraction

The methods of entity extraction include the statistic-based approach and rule-based approach.

Both are adopted in our experiment. Recently, statistic-based approaches of Named Entity Recognition (NER) are widely used. NER labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. It is usually involved in the theories of Maximum Entropy (ME) [4] and Conditional Random Field (CRF) [5]. Here we select a Name Entity Recognizer which is implemented by the Stanford NLP Processing Group [6]. The software provides a general implementation of linear chain CRF sequence models coupled with well-engineered feature extractors for Named Entity Recognition. It provides 3 types of entity, people, organization and location. According to our task, we modify its code and make it output the people and organization. The most serious problem is lack of training data, so we utilize MUC6, MUC7, and ACE et al. to train an appropriate model.

The name of people or organization often appears in some meaningful passage. However, the product may be in a list or table of a product's web site. Only use the sequence labeling approach like CRF is not enough. We also use some rules and manual ways. A part of products and URLs are extracted by generative rules to some special web site. Then we choose them manually.

3.4 Document Retrieval Based Entity

As in eq. 1, the similarity of d and e should be calculated. Some common ways of statistics for all the document and entities are very complex. We adopt the retrieval approach to keep it consistent with the similarity of q and d. Then we can take entities as queries mentioned in the 3.2 section. We get documents related to entities and their similarities. The ones appear in the relevant documents of query are used in the next step. They are just the bridge in the model, see also figure 1. Others are discarded.

3.5 Ranking

After previous work, documents and entities are both received. We need to find the most relative ones by some ranking algorithms. Considering that exponent function is monotonous, and the range is (0,1] while the domain is (-∞,0], it is fit for ranking entities. We define the score of a entity as:

$$Score\left(e_j\right)=\exp\left[-\frac{\alpha}{Sim\left(q,e_j\right)}\right]=\exp\left[-\frac{\alpha}{\sum_{i=1}^{m}Sim\left(q,d_i\right)\cdot Sim\left(d_i,e_j\right)}\right] \tag{2}$$

in which α is a parameter for scaling. For each ranked entity, support documents can be listed by similarity.

Then we allocate a homepage for each entity. In our task, for each target entity at least one homepage must be returned. Documents returned in the homepage and Wikipedia fields must not be retrieved for multiple entities in the same topic. For a topic, we can't ensure that the homepages returned are absolutely correct. If it is forbidden to retrieve for multiple entities in the same topic, once the former homepages are wrong, the later homepages will be influenced. In that condition, for one topic only one entity and homepage (or WP) pair is returned in our experiment this time. So at least a homepage is shown, even some entities have Wikipedia page. HP or WP is determined by:

$$d_{HP,WP}\left(e\right)=\arg\max_{d_i\in D_r}\left\{\exp\left[\frac{sim\left(d_i,e\right)}{\beta}\right]\right\} \tag{3}$$

in which β is a empirical parameter like α .

4. Experiments and Submitted Runs

The Entity Track uses the “Category B” subset of ClueWeb09 dataset which is new. It contains about 50 million English pages. Three quarters are webpage, and others are Wikipedia page. The size decreases from 1.39TB to 267GB by preprocessing.

Table 1. Four runs including numbers of entities and support documents

	Run1(ENT/SP)	Run2(ENT/SP)	Run3(ENT/SP)	Run4(ENT/SP)
Q1	100/18	100/9	100/9	28/12
Q2	100/14	100/15	100/12	5/7
Q3	100/6	89/6	95/6	36/6
Q4	100/6	40/6	100/6	100/20
Q5	7/4	1/1	4/3	2/5
Q6	14/5	24/6	5/4	27/12
Q7	100/6	100/5	98/6	100/16
Q8	11/3	79/6	13/3	12/2
Q9	13/6	11/7	10/5	51/17
Q10	82/6	53/6	88/6	100/20
Q11	51/10	16/11	9/4	44/17
Q12	63/6	100/6	86/6	100/20
Q13	4/3	1/2	3/2	2/6
Q14	21/6	8/5	21/6	100/18
Q15	16/6	26/5	75/6	100/19
Q16	59/6	8/5	4/1	92/19
Q17	4/3	7/5	9/5	18/14
Q18	17/4	65/6	3/5	62/18
Q19	15/6	50/6	23/3	61/19
Q20	54/6	94/7	53/6	100/19

For each query, we return up to 100 related entities. Each entity corresponds to up to 10 supporting documents. We treat different types of query as four runs, original queries, original queries and narratives, terms of queries must appear ordered and retrieving queries in title field. Table 1 lists the 4 runs including numbers of entities and support documents related to 20 topics.

5. Conclusion and Future Work

In our work, an improved two-stage retrieval model is proposed according to the task. The first stage is document retrieval, in order to get the similarity of the query and documents. The second stage is to find the relationship between documents and entities. Final scores are computed by combining previous results. We also focus on entity extraction in the second stage and the final ranking. As of future work, we will care about entity disambiguation and entity-relationship recognition.

Acknowledgements

This work was supported by Project "Multidimensional latent feature extraction model based Evolutional document filtering" supported by NSFC.

References

[1] http://ilps.science.uva.nl/trec-entity/guidelines/

[2] http://www.lemurproject.org/indri/

[3] Zhao Ru. TREC 2005 Enterprise Track Experiments at BUPT. In proceedings of TREC-2005, 2005.

[4] Zhang Suxiang. Automatic Entity Relation Extraction Based on Maximum Entropy. In Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications.

[5] John Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning. 2001.

[6] http://nlp.stanford.edu/